Danso-Appiah et al. address a very important issue: this study aims to provide evidence of the diagnostic performance of POC CCA systematically reviewing and calculating pooled estimates of sensitivity and specificity from available relevant studies to date. This research warrants congratulations.

However, I do have some comments and scepticism re some of the technical parts and conclusions of this study which follow below:

-In page 12 it is explained the problem in helminths’ diagnosis of the lack of a gold standard, i.e. of a diagnostic test of 100 % specificity and 100 % sensitivity. Although I understand the reasoning of creating ‘a parasitological gold standard’, I really cannot understand how we can compare CCA against the combined KK/CCA. In other words, how can we be comparing test A vs testAB. Both tests A and B (i.e. CCA and KK) contain measurement error and in my opinion such a comparison is not very meaningful. I could perhaps see the value of such a comparison if you had a third test C and if you were comparing it to the combined diagnostic test AB.

JM: In principle I consider this form of comparison to be meaningful. The test characteristics, - in terms of sensitivity, specificity, time and cost – of running two diagnostic tests in combination are potentially distinct enough from those of either test in isolation. For example: the decision to use one test may be dependent on the results of the other test, so assumptions of independence between tests cannot be made; the interpretation of test B may be affected by the results of test A; people who conduct both test A and B may be less (or more) adept at applying either test than people who specialise in just one test; there may be economies of scale in using both tests, and so the cost of both tests may be less than the sum of the cost of either test in isolation. For these kinds of reasons we have considered KK/CCA as distinct diagnostic tests in their own right. We have done this when the authors of the primary studies have presented the results for combined tests in this way, implying that the authors of these studies consider the combined KK/CCA option to be distinct.

SUGGESTED ACTION: Add a couple of sentences discussing the above in the report. [All to contribute]

At page 14, for data synthesis and the bivariate metaregression, statistical significance was defined at p-value less than 0.05. Then for assessing heterogeneity, the statistical significance was defined at p-value less than 0.10; I wonder is this not quite an arbitrary choice? Why there are different analysis strategies to assess statistical significance in the same study? Does this not merit at the very least some kind of explanation?

JM: I agree: ultimately the choice of p-values is arbitrary. However, both p value thresholds are very well established conventions for their particular applications. The use of p < 0.05 is long established within the Neyman-Pearson school of hypothesis testing, dating back to approximately the 1920s. However within meta-analysis, which is a much newer subdiscipline of statistics, the paper by DerSimonian & Laird which introduced a test for heterogeneity, and hence that a random effects (RE) rather than fixed effect (FE) model is more appropriate, broke with this tradition in using a more sensitive threshold of p < 0.10 instead. We have simply followed these conventions, by defining heterogeneity as ‘significant’ when p < 0.10, but other hypotheses as ‘significant’ when p < 0.05.

SUGGESTED ACTION: Write something along these lines within the report. [Some advice from Paolo would be useful.]

At page 15, models are not ‘performed’, -statistical models -in general and always- are fitted to data. So I would definitely replace the verb ‘perform’ with ‘fitted’.

ACTION: Make these changes

At page 24, Analysis 4, I am quite sceptical about these results. More precisely, if we are working with a fixed-effect model, then it makes sense to perform a metanalysis as soon as we have two studies, since a summary based on two or more studies yields a more precise estimate of the true effect than either a study alone. But if this is a random effects model as the rest of the analyses, then the estimate of the between-studies variance, will be substantially in error. The standard error of the summary effect is based (in part) on this value, and therefore, if we present a summary effect with confidence interval, not only is the point estimate likely to be wrong but the confidence interval may provide a false sense of assurance. Again my concern about comparing CCA against CCA/KK applies also here. I have the same concern about Analysis 8.

JM: I think I agree. However a corollary of using a small number of studies is that tests for heterogeneity are likely to under-powered too, and so it is difficult to determine, using the Q statistic, whether the FE or RE model is most appropriate, and so FE models could also be misleading. We have performed meta-analyses of subgroups with just two or more studies mainly for completeness, but are also concerned about a risk of false reassurance.

SUGGESTED ACTION: More strongly urge caution in interpreting summary estimates from meta-analyses with (say) less than five studies.

A latent class modelling approach is used for a different purpose (than normally applied in medical research diagnosis) in the current study where it serves as a tool for clustering the studies involved in the meta-analysis of diagnostic accuracy. The 2 latent classes seem to indicate: 1) studies with high sensitivities and high specificities and 2) studies with high sensitivities and low specificities. I am sorry but I failed to understand the use of Table 6. I think this Table merits better explanation. Also I did not understand which studies are in which latent classes. Model 8 -or better a numbered figure would be a better title here- shows 14 studies in latent class 1 and more than 20 studies in latent class 2. Which studies are those? I mean which latent classes contain which studies? Which studies were on the first instance included in the latent class analysis model? Otherwise, what is the point of such an exercise?

[PAOLO TO RESPOND]

Furthermore, in the Discussion section, it is stated that all included studies were cross sectional studies. I am sorry but again I do not agree with this statement. For instance, Koukounari et al. 2013 is a longitudinal study. In Table 1 of Koukounari et al. 2013, it is shown ‘empirical CCA diagnostic performance’ or ‘direct method comparison’ of CCA with 6 KK over 3 days. This table is constructed to specifically show that if one uses 6 KK measurements as a ‘gold standard’ then the CCA diagnostic performance is wrongly estimated over time since KK diagnostic performance varies also over time (or different endemic settings). By using another latent variable modelling approach called Latent Markov modelling (LMM), it was also shown in this same study, that KK diagnostic performance varies over time and days (see Table 2 of this study) while it is explained analytically in the supplementary info how this affects the estimate of the ‘true’ *S. mansoni* prevalence. It is also shown that by using LMM, the diagnostic performance of CCA does not vary over time. I do not think this issue is discussed in the current systematic review.

[PAOLO AND TONY TO DISCUSS AND RESPOND]

I highlight the above issues because I would like again to stress the importance of the validity of gold standard — ideally it should be error free and the medical test under review should be independent of the gold standard (the latter provides support about my concerns with regards of comparing CCA to the combination of CCA/KK mentioned at the begin of this review) as this can increase the area under the curve spuriously. When gold standard is imperfect, sensitivity and specificity of the examined test (in this case CCA) would be affected.

Together with Koukounari et al.2013, there are another 2 more studies:

1. [A five-country evaluation of a point-of-care circulating cathodic antigen urine assay for the prevalence of Schistosoma mansoni.](http://www.ncbi.nlm.nih.gov/pubmed/23339198) Colley DG, Binder S, Campbell C, King CH, Tchuem Tchuenté LA, N'Goran EK, Erko B, Karanja DM, Kabatereine NB, van Lieshout L, Rathbun S. Am J Trop Med Hyg. 2013 Mar;88(3):426-32. doi: 10.4269/ajtmh.12-0639. Epub 2013 Jan 21.
2. [Evaluation of urine CCA assays for detection of Schistosoma mansoni infection in Western Kenya.](http://www.ncbi.nlm.nih.gov/pubmed/21283613) Shane HL, Verani JR, Abudho B, Montgomery SP, Blackstock AJ, Mwinzi PN, Butler SE, Karanja DM, Secor WE. PLoS Negl Trop Dis. 2011 Jan 25;5(1):e951. doi: 10.1371/journal.pntd.0000951.

which used latent class analysis (i.e. assuming not a gold standard) to evaluate CCA diagnostic performance. I personally think it might be worth, to perform a kind of sensitivity analysis and do a metaregression to see if results vary between all other studies and the 3 studies I mention above. Such a comparison could provide some evidence if indeed the use of imperfect gold standards such as KK give wrong estimates of CCA diagnostic performance as highlighted in Koukounari et al 2013 study although I have to recognize that 3 studies might not be enough but definitely still worth to try and see results. Thus, the use of sophisticated statistical modelling that takes into account measurement error I feel warrants some further discussion re recommendations in future research of helminths’ diagnostics.

Finally, I am concerned of how ROC curves are presented. The ROC curve of a test is the graph of the values of sensitivity and specificity that are obtained by varying the positivity threshold across all possible values. What is the point of 7a and 7b analyses?

JM: The ROC curves show how we can expect the joint sensitivity and specificity of the test to vary as we change the diagnostic threshold, IFF we assume our point estimates for the model coefficients are correct. These ROC curves were felt to be more clinically meaningful than the model coefficients themselves, and so more informative to present. They are also required in order to produce AUC estimates, which are often used for at-a-glance comparisons between tests. However we recognise for some of these meta-analyses the assumption above is strong, and so there is potential for overinterpretation.

SUGGESTED ACTION: For tests 7a and 7b, be even clearer that the ROC curves and AUC estimates seem particularly model dependent.

The following:

<http://srdta.cochrane.org/sites/srdta.cochrane.org/files/uploads/Chapter%2010%20-%20Version%201.0.pdf>

should be one of the main citations of this article, in my opinion.

[Tony to respond?]